Untersuchungsdesigns

Klöti, Ulrich ; Widmer, Thomas

mente zu erarbeiten, welche dann unter strategischen Gesichtspunkten umgesetzt werden. Viertens können Evaluationen auch veranlasst werden, um zur *Erhöhung der Akzeptanz* von Massnahmen und zu deren Rechtfertigung beizutragen. In diesem Fall dient Evaluation in erster Linie der Legitimierung. Vor allem bei den letzten beiden Formen besteht die Gefahr einer unzulässigen Instrumentalisierung der Evaluation durch die Politik. Nur die genehmen Resultate werden weiterverwendet (vgl. u.a. Rieder und Varone 1991).

Das Thema "Umsetzung von Evaluationsergebnissen" weist aber noch auf grundsätzlichere Aspekte der Beziehung von Wissenschaft und Politik hin. Im Zentrum steht die Frage, ob wissenschaftliche Beratung tendenziell zu einem Ausschluss demokratischer Kontrolle führt und die Politikumsetzung in die Hand von interessengelenkten Technokraten legt, oder ob Wissenschaft Entscheidungsprozesse offenlegt und Wirkungsüberlegungen anstellt, welche für die Rechtfertigung des Einsatzes von Steuermitteln unerlässlich sind (Knoepfel 1989). Die Dynamiken der Interaktionsprozesse zwischen Wissenschaft und Politik sind insbesondere im Anschluss an die systemtheoretischen Überlegungen von Niklas Luhmann Thema einer noch nicht abgeschlossenen politikwissenschaftlichen Debatte (vgl. Freiburghaus und Zimmermann 1985).

KAPITEL 11: UNTERSUCHUNGSDESIGNS

Ulrich Klöti, Thomas Widmer

Das vorliegende Kapitel befasst sich mit der Wahl der Forschungsstrategie, mit der eine evaluative Frage angegangen wird. Auf die grosse Bedeutung dieses Arbeitsschrittes ist bereits im Kapitel 10.4. hingewiesen worden. Die bei dieser Strategiebestimmung zu treffenden Entscheidungen werden in einem sogenannten Evaluationsdesign zusammengefasst. Darunter verstehen wir die Festlegung der Form der empirischen Umsetzung einer evaluativen Fragestellung. Es geht somit bei der Erstellung eines Evaluationsdesigns darum, die Art und Weise, wie die an die Evaluation gestellte Fragestellung in eine empirische Untersuchungsanlage umzusetzen ist, in kontrollierbarer Weise zu bestimmen. Das Forschungsdesign soll namentlich die Projektmechanik (hypothesengesteuertes Zusammenwirken der zu erklärenden und der erklärenden Variablen) bestimmen, das gewählte Untersuchungskonzept darlegen und begründen sowie die zur konkreten Erfassung der Variablen ausgewählten Indikatoren angeben.

Der Abschnitt 11.1. enthält eine Beschreibung der verschiedenen Design-Dimensionen; diese werden anschliessend verglichen und kritisch diskutiert. Schliesslich werden anhand eines Flussdiagramms die verschiedenen Dimensionen zu einem Ganzen zusammengeführt und damit gleichzeitig eine Synopse der verschiedenen Design-Dimensionen sowie die Interdependenzen zwischen diesen dargestellt.

Im Abschnitt 11.2. geht es darum, anhand von Beispielen einige gängige Evaluationsdesigns vorzustellen, ihre Vor- und Nachteile zu erörtern und ihre Eignung für bestimmte evaluative Fragestellungen zu diskutieren.

Die Fragen der erkenntnistheoretischen Ausrichtung einer Evaluation sowie Beschreibungen und Erörterungen von qualitativen und quantitativen Untersuchungsverfahren finden sich im folgenden Kapitel 12.

11.1. Dimensionen von Evaluationsdesigns

Ein Evaluationsdesign beinhaltet die Umsetzung einer Evaluations-Fragestellung in eine Strategie für die empirische Forschung. Es hat nicht nur der Fragestellung, sondern vor allem auch dem zu evaluierenden Gegenstand (im folgenden Evaluandum genannt) Rechnung zu tragen. Die em-

pirische Sozialforschung hält dazu eine Reihe von Untersuchungskonzepten bereit, die anhand der nachfolgend angeführten zentralen Dimensionen unterschieden werden können:

- Einzelfalluntersuchungen und vergleichende Untersuchungsanlagen,
- Quer- und Längsschnittvergleiche,
- Auswahl und Vollerhebungen sowie
- Experimentelle, quasi-experimentelle und nicht-experimentelle Designs.

In der angeführten Liste fehlt ein Hinweis auf die qualitative bzw. quantitative Untersuchungsmethodik. In unserer Sicht handelt es sich dabei um eine Frage der Forschungs*technik*; sie stellt somit eine verfahrenstechnische Entscheidung dar, die an anderer Stelle diskutiert wird (vgl. Kapitel 12). Sicherlich bestehen zwischen Untersuchungsanlage und Verfahrenstechnik mannigfaltige Interdependenzen. Trotzdem sind zwei unterschiedliche Dimensionen angesprochen, die wir - auch um einen Beitrag zur Versachlichung der Diskussion in diesem Bereich zu leisten - strikt trennen wollen. Entscheidungen auf der Ebene der Untersuchungsanlage bilden keine Präjudizien für die Wahl der Untersuchungstechnik(en). Verfahrenstechnische Fragen haben unseres Erachtens einen ganz klar nachgelagerten Charakter (vgl. dazu Cook und Reichhardt 1979; Bryman 1990).

Etwas anders präsentiert sich die Situation bezüglich der *erkenntnistheoretischen Ausrichtung* von Evaluationen. Auch diese Dimension fehlt in der oben angeführten Liste. Die paradigmatische Ausrichtung einer Evaluationsstudie steht bei der Festlegung des Evaluationsdesigns ebenfalls zur Diskussion. Die Entscheidung über deren Ausgestaltung wird in der Praxis weitgehend bestimmt durch die Ausrichtung der Fragestellung einerseits und die erkenntnistheoretische Position der Evaluatorin oder des Evaluators andererseits. Unseres Erachtens hat sich die erkenntnistheoretische Ausrichtung allerdings der Zielsetzung der Studie, wie sie in der Fragestellung zum Ausdruck kommt, unterzuordnen. Die paradigmatischen Positionen werden im Kapitel 12.1. ausführlicher diskutiert.

11.1.1. Einzelfalluntersuchungen und vergleichende Untersuchungsanlagen

11.1.1.1. Einzelfalluntersuchungen

Die am Einzelfall orientierte Untersuchungsanlage definiert den gesamten Untersuchungsbereich als einen Fall. Es findet kein Vergleich zu anderen Fällen statt. Die Fragestellung, die empirische Untersuchungsanlage, die Evaluationsergebnisse wie auch die darauf beruhenden Empfehlungen sind auf den einen, untersuchten Fall ausgerichtet. Im Grundsatz sind andere Fälle nicht von Interesse. Eine Generalisierung der Evaluationsergebnisse wird weder angestrebt noch ist eine solche Übertragung auf andere Fälle möglich. Denkbar ist höchstens, dass die Adäquatheit der Aussagen, die allenfalls übertragen werden sollten, genauestens überprüft wird, was - da sich die Evaluation auf einen Fall beschränkt - eine weitere Fallstudie erforderlich macht. Die empirisch-analytische Position (vgl. Kapitel 12.1.) lehnt die Generalisierbarkeit von Einzelfalluntersuchungen prinzipiell ab.

Der Ansatz hat den grossen Vorteil, dass der evaluierte Einzelfall detailliert in seiner Gesamtheit untersucht werden kann. Aus der vertieften Analyse eines Falles können für die Praxis relevante Ergebnisse resultieren, und es fallen unter Umständen auch Erkenntnisse von grossem theoretischem Interesse an. Vielfach wird in Abrede gestellt, dass Einzelfalluntersuchungen kausale Wirkungszusammenhänge eruieren könnten. Folgt man den Überlegungen, wie sie von Scriven unter der Bezeichnung "modus operandi" angeführt wurden, ist die Formulierung von kausalen Aussagen aufgrund von Einzelfalluntersuchungen jedoch durchaus möglich (vgl. Scriven 1976 und neuerdings Mohr 1995a: 248-273 und 1995b). In der Evaluationsforschung sind Einzelfalluntersuchungen eher selten; in Tat und Wahrheit kommen bei der Untersuchung der Wirkungen einer Massnahme auch in einem einzigen Fall meist vergleichende Untersuchungsanalysen zur Anwendung, weil meist ein Vorher/Nachher-Vergleich angestellt wird (Längsschnittanalyse).

Generelles Ziel einer Einzelfalluntersuchung bildet das Verständnis oder die Erklärung des Einzelfalls. Dies soll erreicht werden, indem der Fall möglichst detailliert und reich beschrieben wird, um aufgrund dieser Beschreibung weitergehende Aussagen zu ermöglichen (Stake 1995). In der Praxis besteht bei Einzelfalluntersuchungen die Gefahr, dass der deskriptive im Vergleich zum analytischen Aspekt zu viel Gewicht erhält, was den analytischen Gehalt der Evaluation reduzieren kann.

11.1.1.2. Vergleichende Untersuchungsanlagen

Unter dem Sammelbegriff "vergleichende Untersuchungsanlagen" werden verschiedene Konzepte subsumiert, die sich dadurch charakterisieren, dass sie auf einem Vergleich unterschiedlicher Untersuchungsgegenstände aufbauen. Dabei kann die Zahl der Objekte stark variieren. So baut eine sogenannte *Kontrollgruppenuntersuchung* auf der Untersuchung eines Objektes auf, das sodann mit einem oder mehreren (bis auf die zu evaluierende Massnahme oder andere deren Wirkung erklärende Variablen (vgl. Kapitel 5)) möglichst ähnlichen Objekten verglichen wird. Die Vergleichsobjekte werden typischerweise vom Evaluationsgegenstand (Evaluandum) nicht beeinflusst, im Gegensatz zur Untersuchungsgruppe. Aufgrund der beobachteten Differenzen wird dann auf die Wirkungen des zu evaluierenden Programms geschlossen. Eine *"vergleichende Fallstudie"* bezieht verschiedene Fälle gleichgewichtig in die Analyse mit ein. Auch hier wird aufgrund festgestellter Differenzen auf Programmwirkungen geschlossen, wobei bei diesem Untersuchungsansatz Schlüsse zu allen untersuchten Objekten angestrebt werden. Typischerweise führen vergleichende Fallstudien zu Ergebnissen, die sich auf die Gesamtheit eines Falls und nicht auf seine Elemente beziehen. Je nach Vorgehensweise bei der Auswahl der Fälle (vgl. unten Kapitel 11.1.3.) ist es auch möglich, über die konkret untersuchten Fälle hinausgehende Ergebnisse zu erhalten (vgl. Stake 1995: 7-8). Dies setzt jedoch voraus, dass die Auswahl der Fälle genügend breit abgestützt ist. Vielfach sind hier nur hypothetische Aussagen möglich.

Eine weitere Variante vergleichender Untersuchungsanlagen bilden *Längsschnittanalysen*, die einen oder mehrere Fälle über die Zeit hinweg analysieren. Verglichen wird hier zwischen den Zuständen eines Falls zu verschiedenen Zeitpunkten oder in verschiedenen Zeiträumen (vgl. dazu ausführlicher das nachfolgende Kapitel 11.1.2.).

Der letzte Typ von vergleichenden Untersuchungsanlagen, den wir hier erwähnen möchten, ist die *Breitenuntersuchung*. Hier wird eine verhältnismässig grosse Zahl von Vergleichseinheiten dazu verwendet, Rückschlüsse auf die Wirkungen eines Programms zu treffen. Auf eine vertiefte Analyse des Einzelfalls wird hier verzichtet, um eine grössere Zahl von Fällen untersuchen zu können. Weil sich die Untersuchung nur auf einzelne, sehr selektiv ausgewählte Variablen der Fälle bezieht, spricht man dabei auch von Variablenanalysen (Przeworski und Theune 1970, Kriesi 1994: 38 ff.). Gesucht wird in derartigen Studien nach strukturellen

Regelmässigkeiten, die Rückschlüsse auf die Wirkungen des Evaluandums erlauben. Die Analyseeinheit bildet aber auch hier der einzelne Fall, und die Studie nimmt einen Vergleich der Fälle vor.

Die angeführten Typen treten oft in Kombination miteinander auf. Es werden aber auch *Mischformen* eingesetzt (so beispielsweise eine Breitenuntersuchung mit Kontrollgruppenvergleich). Vergleichende Untersuchungsanlagen haben den Vorteil, dass der komparative Ansatz eine erprobte und verhältnismässig aussagekräftige Vorgehensweise darstellt. Bei der Wahl des Typs vergleichender Untersuchungsanlagen ist zu beachten, dass es zwischen interner und externer Validität abzuwägen gilt. Sind generelle Erkenntnisse allgemeiner Art erwünscht (hohe externe Validität), ist eine möglichst grosse Zahl von zu berücksichtigenden Fällen zu wählen. Besteht die Absicht jedoch darin, für den Einzelfall möglichst gut fundierte Ergebnisse zu erhalten (hohe interne Validität), ist die Zahl der Fälle eher klein zu halten (vgl. dazu auch Kapitel 11.2.3.). Beide Eigenschaften (also hohe interne und externe Validität) gleichzeitig anzustreben, verbietet im Regelfall die Forschungsökonomie. Im Rahmen von Evaluationsprojekten ist es kaum möglich, die Vorteile einer Breitenuntersuchung mit jenen einer Einzelfallstudie zu kombinieren. Neben den finanziellen Restriktionen ist dabei auch zu bedenken, dass Ergebnisse von Evaluationsstudien innert nützlicher Frist vorzuliegen haben. Kompromisse sind praktisch immer erforderlich. Die Vor- und Nachteile der verschiedenen Vorgehensweisen sind umsichtig miteinander abzuwägen (vgl. dazu Cook und Campbell 1979: 37-94, besonders 82-85).

11.1.2. Quer- und Längsschnittanalysen

11.1.2.1. Querschnittanalysen

Querschnittanalysen vergleichen verschiedene Untersuchungsobjekte wie Gemeinden, Kantone oder Länder (Makro-Ebene), Organisationen, Verbände, Parteien, Unternehmen usw. (Meso-Ebene) oder Individuen (Mikro-Ebene). Charakteristisch ist, dass die genannten Untersuchungsobjekte zu einem bestimmten Zeitpunkt oder Zeitraum erfasst werden (synchrone Vergleiche). Ein Vorteil von Querschnittanalysen liegt darin, dass der weitere (generelle) Kontext, in dem sich die Objekte lokalisieren lassen, für alle Untersuchungsobjekte derselbe ist. So sieht sich eine privatwirtschaftliche Unternehmung im Kanton A den gleichen weltwirt-

schaftlichen Herausforderungen ausgesetzt, wie dies für ein Unternehmen im Kanton B der Fall ist. An diesem Beispiel lässt sich aber auch eine Problematik von Querschnittanalysen aufzeigen. Die Vergleichbarkeit der Untersuchungsobjekte ist oft nur schwer herzustellen. So muss beim Vergleich der beiden Unternehmen im obigen Beispiel Gewähr dafür bestehen, dass sie in den gleichen Märkten tätig sind, eine ähnliche Kostenstruktur und eine vergleichbare Grösse aufweisen, usw. Um diese Vergleichbarkeit herzustellen, bestehen verschiedene Ansätze. So können die Gruppen derart aus Objekten zusammengestellt werden, sodass ein Objekt aus Gruppe A mit einem Objekt aus Gruppe B ein möglichst ähnliches (Vergleichs-)Paar bildet. Eine andere Möglichkeit liegt darin, dass die Objekte nach dem Zufallsprinzip ausgewählt werden. Dies setzt jedoch voraus, dass eine genügend grosse Zahl von Vergleichseinheiten (Gesetz der grossen Zahl) und keine strukturellen Differenzen zwischen den zwei Gruppen bestehen. Weiter kann versucht werden, möglicherweise relevante Drittfaktoren zu kontrollieren.

Ein Vorteil von Querschnittanalysen bildet ihre verhältnismässig gute Handhabbarkeit. Im Gegensatz zu Längsschnittvergleichen sind Querschnittstudien methodisch einfacher durchzuführen und haben den in der Evaluationsforschung bedeutenden Vorteil, dass Aussagen aufgrund einer Einmal-Erhebung bereits möglich sind. Dies erlaubt, die Evaluationsergebnisse rechtzeitig vermitteln zu können. Im Gegensatz dazu eignen sich jedoch Querschnittanalysen weniger dazu, dynamische Prozesse zu erfassen. Sie bauen auf einem statischen Konzept auf. Deshalb empfiehlt sich bei der Untersuchung von stark volatilen Gegenständen der Beizug von Längsschnittansätzen.

Interkantonale und insbesondere internationale Querschnittvergleiche werden oft von unterschiedlichen Forschungsteams durchgeführt (Reisekosten, Ortskenntnisse etc.). Dabei besteht erfahrungsgemäss die Gefahr, dass bei den Fallstudien unterschiedliche Variablen und Indikatoren erhoben werden. In solchen Projekten ist eine genaue Verständigung über das gemeinsame Forschungsdesign und gegebenenfalls eine permanente Kontrolle über dessen Einhaltung von grosser Bedeutung (Weidner und Knoepfel 1983).

11.1.2.2. Längsschnittanalysen

Bei Längsschnittanalysen wird derselbe Gegenstand über die Zeit hinweg untersucht. Seine Entwicklung wird über genau identifizierbare Phasen

hinweg beobachtet (diachrone Vergleiche). In der Evaluationsforschung wird dabei typischerweise ein Gegenstand vor und nach der Einführung des zu evualuierenden Programms analysiert, um aufgrund auftretender Veränderungen auf Programmwirkungen zu schliessen. Aber nicht nur die Einführung eines Programms, auch Veränderungen an bestehenden Programmen oder Programmbeendigungen ("program termination", vgl. Brewer 1978; deLeon 1983 und 1987) können in dieser Form analysiert werden (vgl. Benninghoff 1995). In der Evaluationsforschung werden derartige Neuerungen als Interventionen bezeichnet. Da sich die Evaluationsforschung oft mit Veränderungshypothesen befasst, erscheinen solche Längsschnittdesigns oft als die adäquateren Ansätze als diachrone Untersuchungsanlagen; denn gesellschaftliche Veränderungen beanspruchen üblicherweise Zeit.

Längsschnittanalysen können in unterschiedlicher Weise ausgestaltet werden. Als klassisch zu bezeichnen ist das sogenannte *"pretest-posttest design"*, bei dem der Zustand zu einem Erhebungszeitpunkt vor der Intervention mit jenem nach dieser verglichen wird. Im Regelfall werden dabei die Daten jeweils trotzdem nur in einer einzigen Feldbegehung erhoben. Es ist aber durchaus auch möglich, den Status zu mehreren Zeitpunkten vor und nach der Intervention zu erfassen. Dehnt man die Datenerhebung weiter aus, so dass sowohl in der Prä- wie auch in der Postinterventionsphase eine Vielzahl von Beobachtungen zur Verfügung stehen, spricht man von einer *Zeitreihenanalyse*. Mit dieser Untersuchungsanlage lässt sich eine höhere Verlässlichkeit der Ergebnisse erreichen, als dies bei einem einfachen "pretest-posttest design" der Fall ist (Widmer 1991: 123-125). Hingegen ist es vielfach aus praktischen Überlegungen nicht möglich, ein Zeitreihendesign umzusetzen, da Ressourcen zur Datenerhebung fehlen oder eine längerfristige Datenerhebung nicht möglich ist. Im Gegensatz dazu stellt das "pretest-posttest design" oft eine recht praktikable Lösung dar.

11.1.2.3. Kombination und Integration

Wie aus dem bereits Gesagten hervorgeht, haben sowohl Längs- wie Querschnittanalysen ihre spezifischen Vor- und Nachteile. Um die Leistungsfähigkeit der Designs zu steigern, bietet sich die Möglichkeit an, eine Kombination der beiden Untersuchungsanlagen zu wählen (Knoepfel 1995e: 49 ff.). Dies kann einerseits dadurch geschehen, dass zwei getrennte Analysen durchgeführt werden. Dadurch lassen sich die Analyse-

ergebnisse gegenseitig ergänzen, besser fundieren oder auch überprüfen. Hier spricht man von einer *Kombination* der Untersuchungsanlagen. Fliessen beide Ansätze in ein einziges Design ein, nennt man dies *Integration*.

Beide Vorgehensweisen haben, im Gegensatz zu der ausschliesslichen Verwendung eines Längs- oder Querschnittdesigns, den grossen Vorteil, dass sie in der Lage sind, gleichermassen statische Phänomene wie auch dynamische Prozesse zu erfassen. Die dafür erforderlichen Aufwendungen verhindern jedoch in der Praxis der Evaluationsforschung häufig deren Verwendung (vgl. dazu Cook 1985; Denzin 1989: 234-247; Patton 1990: 464-472; Hakim 1992: 144-145).

11.1.3. Auswahl und Vollerhebung

Evaluationen können sich mit einer Auswahl oder mit allen bestehenden Objekten befassen.

11.1.3.1. Auswahl

Bei Evaluationen, die nur eine bestimmte Auswahl (auch als Stichprobe oder "sample" bezeichnet) untersuchen, sind vor allem zwei Punkte wichtig: die Vorgehensweise bei der Auswahl und die Gültigkeit der Ergebnisse.

Erstens besteht das Problem, *wie die Auswahl getroffen* werden soll, also wie aus der Menge aller Objekte (Grundgesamtheit oder auch Universum genannt) jene ausgewählt werden sollen, die in der Evaluation empirisch untersucht werden (vgl. dazu Henry 1990). Dabei sind grundsätzlich zwei verschiedene, systematische Vorgehensweisen zu unterscheiden. Die Wahrscheinlichkeits- oder Zufallsauswahl ("random sample") baut darauf auf, dass für alle Elemente der Grundgesamtheit dieselben Chancen bestehen, in die Stichprobe aufgenommen zu werden. Eine gezielte Auswahl ("purposive sample", "theoretical sample") versucht dagegen, aufgrund bestimmter Dimensionen die Eigenschaften der Elemente in der Grundgesamtheit in systematischer Weise in der Auswahl abzubilden. Derartige Stichproben sollen beispielsweise hinsichtlich gewisser Eigenschaften möglichst homogen oder heterogen sein. Auch das sogenannte Schneeballverfahren, bei dem ein Untersuchungsobjekt auf weitere Untersuchungsobjekte hinweist, lässt sich dieser Kategorie zuordnen. Die Vorgehensweise beim Treffen einer Auswahl hat massgeblichen Einfluss auf die Aussagekraft einer Evaluation (für eine ausführlichere Diskussion siehe Ackoff 1953: 123-126).

Es ist daher nötig, das Design vor der Auswahl der Untersuchungseinheiten genau festzulegen und darin zu bestimmen, bei welchen Variablen Konstanz bzw. Varianz gesucht wird. Ausserdem ist es sinnvoll, in einer als Vollerhebung konzipierten Vorstudie ("Survey") einige leicht erhebbare Merkmale aller möglichen Untersuchungseinheiten der Grundgesamtheit zu analysieren. Vielfach ist eine solche Survey-Studie für die - spätere - Interpretation der Ergebnisse und für die genaue Situierung des Untersuchungsgegenstandes ebenso hilfreich wie für die Bereitstellung einer Argumentation gegenüber allfälligen Kritiken einer vergleichenden Evaluation (Argument des unzulässigen Vergleichs).

Zweitens entsteht bei Evaluationen die Frage, inwiefern die Untersuchungsergebnisse auf Gegenstände ausserhalb der empirisch untersuchten Objekte *übertragen* werden können (Generalisierung). Dies hängt wiederum sehr stark davon ab, wie die Stichprobe gewonnen wurde. Gehen wir von einer Wahrscheinlichkeitsauswahl aus, sind Schlüsse auf die nicht in der Stichprobe enthaltenen Elemente der Grundgesamtheit zulässig, sofern die Zahl der untersuchten Objekte ausreichend gross ist. Bei einer gezielten Auswahl ist die Generalisierung der Untersuchungsergebnisse auf theoretischer Ebene zu belegen. Insbesondere muss überzeugend nachgewiesen werden können, dass die zur Stichprobenziehung berücksichtigten theoretischen Dimensionen relevant sind und damit die Stichprobe alle für die Evaluation wichtigen Merkmale der Grundgesamtheit in adäquater Form abbilden.

Generell ist festzuhalten, dass empirische Aussagen nie auf Gegenstände übertragen werden können, die nicht Bestandteil der Grundgesamtheit sind. Generalisierungen dieser Art bedürfen weiterer Abklärungen.

11.1.3.2. Vollerhebung

Bei einer Vollerhebung werden alle Elemente einer definierten Grundgesamtheit in die empirische Untersuchungsanlage miteinbezogen. Grundgesamtheit und Stichprobe fallen damit zusammen. Die beim Treffen einer Auswahl relevanten kritischen Aspekte (Art der Stichprobenziehung und Verallgemeinerung der Resultate) entfallen. Trotzdem werden in der Evaluationsforschung häufig keine oder nur grobe Vollerhebungen (für die Fallauswahl) durchgeführt. Das liegt daran, dass normalerweise eine relativ grosse Zahl von Elementen in der Grundgesamtheit existieren. Der

vollumfängliche Einbezug aller Elemente würde vielfach den Rahmen der Evaluation sprengen; der dafür notwendige Aufwand könnte kaum gerechtfertigt werden.

Dennoch gibt es einige Situationen, bei denen eine Vollerhebung eingesetzt wird. An erster Stelle ist auf den Fall hinzuweisen, bei dem die Grundgesamtheit gerade aus einem Fall besteht. Hier wird gezwungenermassen eine Vollerhebung durchgeführt. Aber auch in Situationen, in denen die Grundgesamtheit nur aus einigen wenigen Fällen besteht, wird häufig die Vollerhebung gewählt. Je nach Standardisierungsgrad der Erhebungs- und Analyseinstrumente kann auch bei einer grösseren Zahl von Elementen noch eine Vollerhebung ins Auge gefasst werden.

11.1.4. Experimentelle, quasi-experimentelle und nicht-experimentelle Designs

Wie weit ein Design als experimentell bezeichnet wird, hängt von der Vorgehensweise bei der Bestimmung der Vergleichsgruppe und der Art der Verteilung der Untersuchungsobjekte auf die Vergleichsgruppen ab. Hinsichtlich der Hypothese stellt die Bildung der Vergleichsgruppen im Prinzip die Zuordnung der Untersuchungsobjekte zu den Ausprägungen der unabhängigen oder der abhängigen Variable dar. In der Evaluationsforschung werden also beispielsweise zwei Vergleichsgruppen gebildet, wovon die erste aus Individuen besteht, die an dem zu evaluierenden Programm teilnehmen (Versuchsgruppe, "experimental group"), während die zweite Gruppe (Kontrollgruppe, Vergleichsgruppe, "comparison group") dies nicht tut. Aufteilungen dieser Art können vor (ex ante-Design) oder nach der Datenerhebung (ex post facto-Design) erfolgen. Wesentlich ist in allen Fällen, dass sich die beiden Gruppen - bis auf die Betroffenheit durch das Evaluandum oder das Explanandum - möglichst ähnlich sind (vgl. dazu und zum folgenden Campbell und Stanley 1963; Cook und Campbell 1979 und Mohr 1995a).

11.1.4.1. Experimentelle Designs

Experimentelle Untersuchungsanlagen gehen von der Überlegung aus, dass aufgrund einer Manipulation der unabhängigen Variablen - in der Evaluationsforschung also des Evaluandums - der Wirkungsnachweis erbracht werden kann, also kausale Zusammenhänge isoliert werden können. Grundsätzlich sind zwei experimentelle Untersuchungsanlagen zu differenzieren: das Laborexperiment und das Feldexperiment. Während sich das Feldexperiment in einem "natürlichen" Kontext abspielt, wird beim Laborexperiment das Umfeld künstlich hergestellt. Experimente gehen davon aus, dass die oben beschriebene Zuordnung der Objekte zu den Untersuchungsgruppen durch eine Zufallsauswahl (sogenannte "Randomisierung") vor der Datenerhebung erfolgt. Diese Bedingung ist jedoch aus ethischen, politischen oder sozialen Gründen sehr oft nicht gegeben. So ist beispielsweise die Vergabe von Subventionen an zufällig ausgewählte Empfänger - während andere Personen in identischen Situationen diese nicht erhalten - nicht zulässig. Aus diesen Gründen sind echte Experimente mit Zufallsauswahl in der Praxis der Evaluation von öffentlichen Politiken nur selten anzutreffen.

11.1.4.2. Quasi-experimentelle Designs

Quasi-Experimente folgen - wie die echten Experimente - grundsätzlich derselben Logik, wobei die Anforderungen an Experimente (ex ante-Design, Randomisierung) nicht vollständig einzuhalten sind. Quasi-Experimente versuchen aber der experimentellen Vorgehensweise möglichst nahe zu kommen. Dazu behilft man sich verschiedener Instrumente, um den Verstoss gegen die experimentellen Anforderungen (zumeist die Randomisierung) teilweise auszugleichen. Häufig ist dabei das sogenannte paarweise Matching. Angestrebt wird hier, dass sich zu einem Untersuchungsobjekt in der Versuchsgruppe jeweils ein hinsichtlich allen relevanten Drittfaktoren möglichst ähnliches Objekt in der Vergleichsgruppe findet. Eine andere Möglichkeit besteht darin, Drittvariablen im Nachhinein durch die Verwendung von multivariaten, statistischen Verfahren zu kontrollieren (vgl. dazu unten Kapitel 12.3.). Diese Instrumente sind nicht in der Lage, die fehlende Randomisierung wettzumachen. Da aber sehr häufig ein echtes Experiment nicht realisiert werden kann, stellen Quasi-Experimente eine valable Alternative dar (Hedrick, Bickman und Rog 1993: 58).

11.1.4.3. Nicht-experimentelle Designs

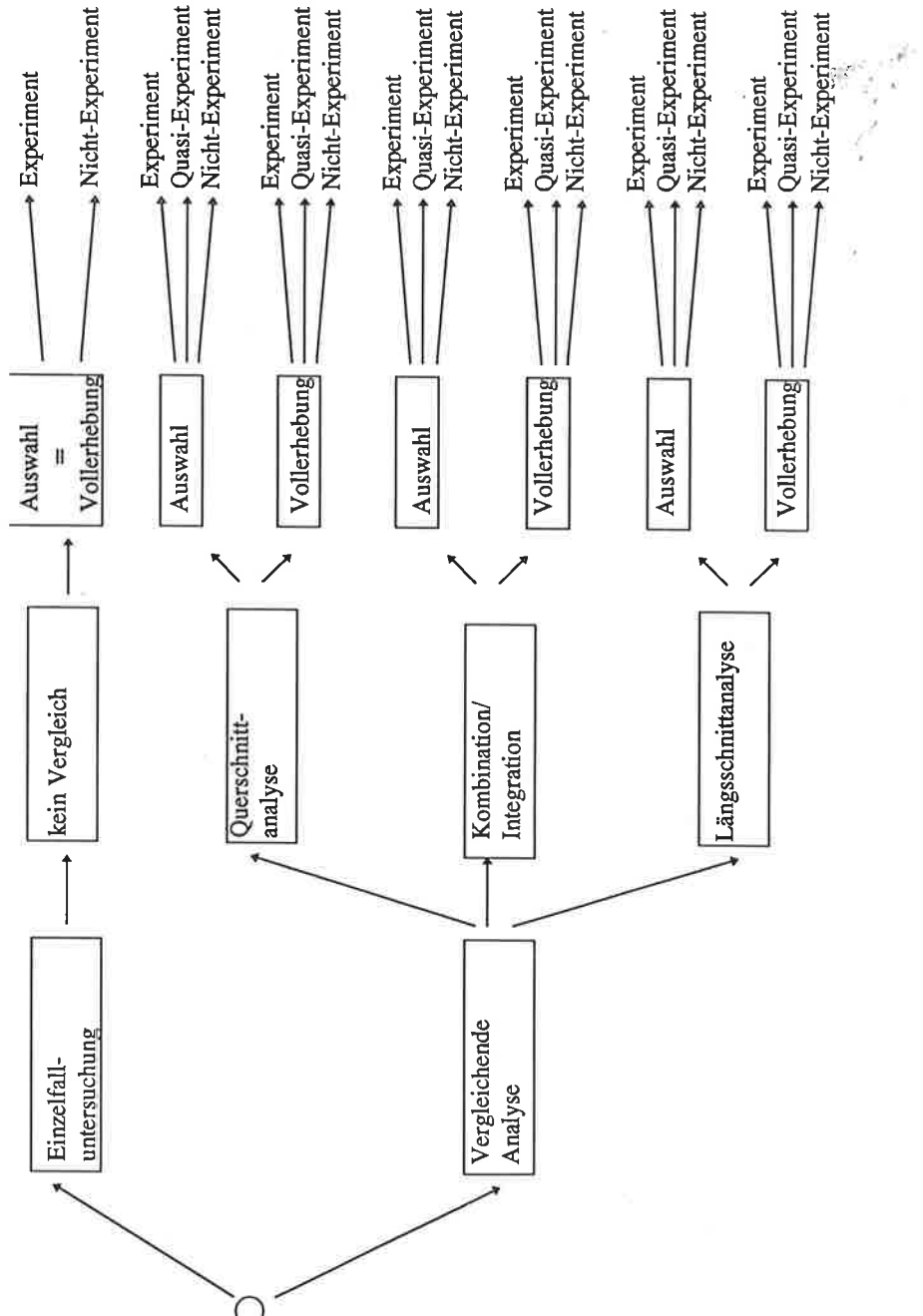
Unter der Bezeichnung der nicht-experimentellen Designs verstehen wir verschiedene Untersuchungsanlagen, die sich nicht auf eine experimentelle Logik stützen. Diese werden teilweise auch als "vorexperimentelle

Designs" bezeichnet, was darauf hinweisen soll, dass diese Untersuchungsanlagen im Rahmen der experimentellen Denkweise als wenig leistungsfähig eingestuft werden. Nicht-experimentelle Designs sind in der Evaluationsforschung relativ weit verbreitet, sind sie doch im allgemeinen die am einfachsten zu implementierende Designform. Beispiele sind etwa Evaluationen, die auf einer "One-shot case study" basieren, oder die klassischen ex post facto-Versuchsanlagen. Während Vertreter der experimentellen Schule diese als nichtaussagekräftig zurückweisen, wird von Seiten anderer Methodiker auf die spezifische Leistungsfähigkeit dieser Untersuchungsanlagen verwiesen. Insbesondere werden dabei die Möglichkeiten zur Hypothesengenerierung (im Gegensatz zur Absicht der Hypothesenüberprüfung bei experimentellen und quasi-experimentellen Designs), aber auch die Einfachheit der Untersuchungsanlage hervorgehoben. Die andere Stossrichtung der nicht-experimentellen Untersuchungsanlagen kommt auch darin zum Ausdruck, dass sie auch als deskriptive Designs bezeichnet werden (siehe etwa Hedrick, Bickman und Rog 1993: 44-51).

11.1.5. Entscheidungsbaum zur Ausgestaltung eines Evaluationsdesigns

Die vier hier diskutierten Dimensionen zur Definition eines Untersuchungsdesigns lassen sich recht weitgehend miteinander kombinieren, auch wenn in der Praxis gewisse Kombinationen sehr häufig oder nur sehr selten eingesetzt werden. Die nachfolgende Graphik soll dazu dienen, die Kombinationsmöglichkeiten in den unterschiedlichen Dimensionen zu veranschaulichen. Dabei haben wir die Darstellungsform eines Entscheidungsbaums gewählt, obwohl dies den falschen Eindruck erwecken könnte, dass diese Entscheide nacheinander in einer bestimmten Reihenfolge zu treffen wären. Dies ist aber in keiner Weise zutreffend. Die Definition eines Untersuchungsdesigns ist ein iterativer Prozess, bei dem keine bestimmte Reihenfolge einzuhalten ist und bei dem auf getroffene Entscheide (zumindest solange sich die Evaluation noch in der Planungsphase befindet) immer wieder zurückgekommen werden kann und oft auch soll.

Abbildung 4: Entscheidungsbaum zur Definition eines Untersuchungsdesigns



Aus der Darstellung wird deutlich, dass bestimmte Kombinationen der unterschiedlichen Dimensionen nicht denkbar sind. Dies gilt besonders im Bereich der Einzelfalluntersuchungen, bei denen Auswahl und Vollerhebung identisch sind und keine quasi-experimentellen Untersuchungsanlagen bekannt sind.

Im nachfolgenden Kapitel 11.2. werden nun die hier diskutierten Designformen anhand von typischen Kombinationsmöglichkeiten vertieft behandelt und mit Beispielen aus der Evaluationspraxis erläutert. Ebenso wird in diesem Abschnitt zu diskutieren sein, welche Designtypen in der Evaluationspraxis besonders häufig respektive besonders selten eingesetzt werden und welche Gründe für diese Differenzen verantwortlich sind.

11.2. Häufig verwendete Untersuchungsdesigns

In der Evaluationspraxis ist es nicht immer möglich, eine in jeder Hinsicht optimale Untersuchungsanlage zu wählen. Aus Gründen der sicheren kausalen Zuordnung von Wirkungen zu bestimmten Programmen wäre es zwar z.B. wünschbar, systematisch angelegte Experimente mit einer Vollerhebung über längere Zeiträume durchzuführen. Aus forschungsökonomischen Erwägungen, weil für frühere Zeitpunkte keine Daten vorliegen oder erhoben werden können und wegen des Erfordernisses, eine Evaluationsstudie rasch abschliessen zu müssen, sehen sich Evaluationsteams indessen häufig veranlasst, auf aufwendige Designs zu verzichten und sich auf Momentaufnahmen eines Einzelfalles zu beschränken.

In diesem Optimierungskalkül, dem letztlich alle Evaluationen unterworfen sind, schälen sich immer wieder ähnliche *Typen* von Evaluationsdesigns als gangbare Lösungen heraus. Einige dieser häufig verwendeten Designs sollen in der Folge umschrieben, anhand von Beispielen erläutert und auf ihre Vor- und Nachteile hin untersucht werden. Dabei sollen folgende Typen zur Sprache kommen:

- Die Einzelfallstudie
- Der Fallstudienvergleich
- Der quasi-experimentelle Vorher/Nachher-Vergleich
- Die Zeitreihenanalyse
- Die Kombination von Design-Ansätzen.

Der Beschreibung der genannten Evaluationstypen ist vorauszuschicken, dass sich die einzelnen häufig verwendeten Designs nie in völlig reiner Form auf die in Grafik 1 unterschiedenen Dimensionen umlegen lassen. Vielmehr handelt es sich in der Praxis oft um Mischformen, die z.B. einen Querschnitt mit einem Längsschnitt verbinden, die nur teilweise experimentellen Charakter haben und die für gewisse Aspekte der Evaluation eine Vollerhebung wählen, für andere Aspekte indessen lediglich eine Auswahl der Untersuchungsobjekte analysieren. Diese Mischung von Elementen in einzelnen Untersuchungsdesigns wird nicht immer bewusst vorgenommen und schon gar nicht immer explizit gemacht. Dass sie aber vorkommt, braucht nicht weiter zu verwundern. Nicht selten wird nämlich ein gemischtes Untersuchungsdesign durch die Sachlage, durch Wünsche der auftraggebenden Stelle, durch begrenzte Mittel für die Evaluation und durch die Daten- und Quellenlage weitgehend vorgegeben.

11.2.1. Die Einzelfallstudie

Bei der Einzelfallstudie gilt der gesamte Untersuchungsbereich als einziger Fall. Dieser wird gesamtheitlich betrachtet. Der Fall, der für die Evaluation besonders interessante Aspekte aufweisen sollte, wird auf möglichst viele Dimensionen hin beobachtet, beschrieben und untersucht. Damit soll verhindert werden, dass ein Untersuchungsobjekt auf einige wenige Variablen reduziert wird. Dies würde dem Untersuchungssubjekt nicht gerecht, weil das Herausgreifen einzelner Merkmale einigermassen willkürlich erscheint, die Komplexität des Gegenstandes verkürzt und die Individualität und Identität des zu Untersuchenden verletzten könnte. Tatsächlich geht es darum, ein ganzheitliches und damit realistisches Bild der sozialen Welt zu zeichnen. Mithin sind möglichst alle für das Untersuchungsobjekt relevanten Dimensionen in die Analyse einzubeziehen (Lamnek 1989: 5).

Mit dieser Umschreibung ist angedeutet, dass Einzelfallstudien häufig nicht von einem empirisch-analytischen, sondern von einem hermeneutischen Forschungsparadigma ausgehen (vgl. 12.1.). "Die Einzelfallstudie im qualitativen Paradigma strebt eine wissenschaftliche Rekonstruktion von Handlungsmustern auf der Grundlage von alltagsweltlichen, realen Handlungsfiguren an. Dabei versucht der Forscher nicht nur als alltagsweltlicher Handlungspartner die Figuren nachzuvollziehen, sondern diese in den wissenschaftlichen Diskurs zu überführen und Handlungsmuster zu

identifizieren, indem er allgemeinere Regelmässigkeiten vermutet" (Lamnek 1989: 16).

Die Einzelfallstudie verzichtet im Prinzip auf irgendeinen Vergleich. Sie orientiert sich weder an früheren Situationen (Längsschnitt), an anderen Fällen (Querschnitt) noch an einer vorgegebenen oder von den Forschern selbst entwickelten Sollgrösse. Wie bereits erwähnt (Kapitel 11.1.1.), ziehen freilich auch Einzelfallstudien oft implizit Vergleiche. So machen sie etwa die Aussage, auch nach der Einführung eines Arbeitsbeschaffungsprogramms seien in einer Region viele Arbeitslose anzutreffen. Sie meinen damit mehr Arbeitslose als früher, mehr als erwartet (Soll-Grösse) oder mehr als in anderen Regionen.

In der grossen Mehrheit arbeiten Einzelfallstudien mit qualitativen Forschungstechniken (vgl. 12.2.). Sie bevorzugen offene oder wenig strukturierte Interviews mit wenigen Betroffenen oder Fachleuten, interpretieren Dokumente und verzichten weitgehend auf quantitative Indikatoren. Angesichts ihrer methodischen Ausrichtung erhalten Einzelfallstudien häufig einen stark deskriptiven Anstrich.

Einzelfallstudien können allerdings durchaus grossen analytischen Gehalt haben. In welchem Ausmass dies der Fall ist, hängt stark von der Auswahl des untersuchten Falles ab. In einer Anleitung des GAO (1990) zur Durchführung von Fallstudien nimmt dieser Arbeitsschritt deshalb breiten Raum ein. Als Oberkriterien werden Bequemlichkeit, Zufallsauswahl und erwarteter Nutzen unterschieden. Das Bequemlichkeitskriterium kommt dann zum Zug, wenn jener Fall untersucht werden soll, über den am leichtesten Informationen erhältlich sind. Diesem pragmatischen Vorgehen steht die reine Zufallsauswahl gegenüber.

Nach unserer Auffassung ist allerdings die Auswahl des Falles streng an den Zielen der Evaluation zu orientieren. Das GAO unterscheidet dabei sieben verschiedene Selektionskriterien. Erstens ist es möglich, ein *Extrem* auszuwählen, um dadurch Hinweise auf die Ursachen für grosse Varianzen zu bekommen. Als zweites sind *Musterbeispiele* denkbar. Sie legen dar, was zum Erfolg in einzelnen Fällen beigetragen hat. Das Gegenbeispiel stellt der *kritische Fall* dar. Er weist auf spezifische Probleme hin. Weiter ist es möglich, einen *typischen Fall* auszuwählen. Er gibt Aufschluss darüber, wie die Wirkungskette im "Normalfall" zusammenhängt. Durch die Beschreibung von Fällen, welche ein spezielles Cluster vertreten, lassen sich verschiedene Programmtypen vergleichen. Ein *illustrativer Fall* kann dazu dienen, Probleme oder Chancen von Programmen besonders eindrücklich darzulegen. Als letzten Typ bezeichnet das GAO

Spezialfälle. Dabei wird ein herausragendes Ereignis untersucht. Dabei kann es sich beispielsweise um einen Fall handeln, der besonders viel Aufsehen erregt hat. Für eine solche Situierung des Falles sind, wie erwähnt (Kapitel 11.1.1.), vorgängige Survey-Studien hilfreich.

Ein gutes *Beispiel* für eine Einzelfallstudie ist die Arbeit von Michel Rey und seiner Equipe (Rey et al., 1993) zur *Wirtschaftsförderungspolitik des Kantons Wallis*. Die Autoren gehen vom Anreizcharakter dieses politischen Programms aus und leiten daraus ab, dass die Wirtschaftsförderung nur dann eine Wirkung entfalten könne, wenn sie eine Veränderung des Verhaltens der Adressaten auslöste. Dazu müssten die Adressaten die Politik nicht nur kennen und verstehen, sondern sie müssten auch eine positive Meinung darüber haben.

Die Evaluatoren führten deshalb mit rund 60 Walliser Akteuren Gespräche über verschiedene Fragen, die von einer zwölfköpfigen Evaluationsinstanz als problematisch bezeichnet worden waren. Aus den dabei gewonnenen qualitativen Aussagen haben die mit der Evaluation Beauftragten, die sich selbst als Experten bezeichnen, Schlussfolgerungen gezogen und Empfehlungen zuhanden der für die Wirtschaftsförderungspolitik Verantwortlichen gezogen. Angesichts der umstrittenen und enttäuschenden Ergebnisse sind im folgenden zwei Arbeitsgruppen eingesetzt worden, die in einem iterativen Prozess neue Ziele, Strategien, Vollzugs- und Koordinationsverfahren zu entwickeln hatten. Dieser Prozess der Weiterentwicklung der Politik sei fortzusetzen, schliessen die Autoren.

Das Beispiel zeigt die Vor- und Nachteile von derart konzipierten Einzelfallstudien auf. Als Vorteil ist der Umstand zu werten, dass ohne Vorurteile die verschiedenen möglichen Gründe für den (Miss-)Erfolg der Politik ermittelt werden konnten. Das Vorgehen bot auch die Möglichkeit, flexibel auf neue Situationen zu reagieren, Erhebungsdimensionen zu verändern und eine zweite Phase anzuschliessen. Das Untersuchungsdesign erwies sich auch deshalb als angebracht, weil die Wirkungszusammenhänge noch wenig bekannt waren und explorativ neue Zusammenhänge ermittelt werden konnten.

Umgekehrt ist das Vorgehen auch mit beträchtlichen Risiken verbunden. Die Verlässlichkeit der aufgezeigten Zusammenhänge bleibt teilweise ungesichert. Die Qualität der Resultate hängt stark von der sozialwissenschaftlichen (und sozialen!) Kompetenz des Evaluationsteams ab. Die Aussagen der Adressaten garantieren nicht, dass bei einer Änderung der Politik eine grössere Wirksamkeit zu erwarten ist. Eine Übertragbarkeit auf andere Fälle ist genau so wenig möglich, wie in anderen Kantonen mit

der entsprechenden Politik gewonnene Erfahrungen für den Untersuchungsfall genutzt werden können. Das Vorgehen ist somit aufwendig und in seinem Erfolg unsicher.

11.2.2. Vergleichende Fallstudien

Mit dem Vergleich mehrerer (wenigstens zweier) Fallstudien sollen die wichtigsten Nachteile der Einzelfallstudie behoben werden. Das Vorgehen bei der Erstellung der einzelnen Fallstudien ist in grossen Zügen dasselbe wie bei einer Einzelfallstudie. Im Hinblick auf die im Kapitel 11.1. angeführten Dimensionen ist zudem zu sagen, dass es sich auch bei den vergleichenden Fallstudien nicht um eine Art Experiment handelt und dass keine Vollerhebung angestrebt wird. Es geht im übrigen fast immer um Querschnittanalysen. Es ist allerdings auch hier zu betonen, dass die genannte Kombination von Merkmalen in den vier Design-Dimensionen keineswegs zwingend, wohl aber typisch ist und dass das Design der im Querschnitt vergleichenden Fallstudien auch nicht immer in seiner reinen Form zur Anwendung gelangt.

Solche vergleichend angelegte Fallstudien sind wohl das in der schweizerischen Evaluationsforschung am häufigsten verwandte Design. Es ist deshalb nicht ganz einfach, aus der Fülle der vorhandenen *Beispiele* das geeigneteste auszuwählen. Solche Studien aus jüngster Zeit finden sich bei Terribilini (1995)[1], Gerheuser und Schmid (1993)[2], Knoepfel, Imhof und Zimmermann (1995)[3], Knoepfel, Kissling-Näf und Marek (1997)[4] oder Klöti, Haldemann und Schenkel (1993)[5]. Angesichts ihres hermeneutisch und qualitativ orientierten Vorgehens haben wir uns für die Arbeit von Willy Bierter und Hans-Martin Binder (1993) zur *wirtschaftlichen Innovationsförderung* entschieden. Die beiden Autoren gehen in dieser Untersuchung der Frage nach, wie weit staatliche Förderungsleistungen für den unternehmerischen Innovationsprozess wirksam sind. Dabei analysieren sie die Wirkungen der Instrumente des Bundes und der Kantone Solothurn und St. Gallen. Diese Auswahl der Fälle wird damit begründet, dass sie

1 Verkehrsberuhigungsmassnahmen in ein einigen Schweizer Städten.
2 Lohngleicheitspostulat in der Heimarbeit in verschiedenene Kantonen.
3 Massnahmen verkehrsbezogener Umweltpolitik.
4 Evaluation von Lernprozessen aus der Umwelt- und Gesundheitspolitik anhand von 28 Fallstudien.
5 Vergleich der Umwelt- und Verkehrspolitik einiger schweizerischer Grossstädte.

Varianz ermögliche, und zwar "sowohl auf der Ebene unterschiedlicher "Philosophien" von Wirtschafts- und Innovationsförderungspolitik hinsichtlich Zielperspektiven, Strategien und Mitteleinsatz, als auch auf der Ebene unterschiedlicher Struktur-Merkmale (Wirtschaftsstruktur, regionales Umfeld usw.)" (Bierter und Binder 1993: 17). Die Studie stützt sich auf rund 40 Gespräche mit Vertretern von Bundesämtern, kantonalen Verwaltungen, Banken, betroffenen Unternehmungen, Verbänden sowie Forschungs- und Entwicklungsinstitutionen. Einer Runde von qualitativen Interviews folgten in einer zweiten Phase Gruppengespräche zur kritischen Überprüfung der ersten Befunde. Damit wird auch die Zeitdimension in die Untersuchung eingeführt, werden doch die in der späteren Phase von den früheren Aussagen abweichenden Haltungen der Akteure wenigstens teilweise auf die veränderten konjunkturellen und politischen Rahmenbedingungen zurückgeführt.

Das Vorgehen hat den Vorteil, dass es eine tiefe Durchdringung der drei Fälle erlaubt. Bei der Auswertung und bei den Schlussfolgerungen sind die Möglichkeiten, die der Vergleich bot, freilich nicht voll genutzt worden, da in den drei Fällen kein systematischer Zusammenhang zwischen den unterschiedlichen Ausgangslagen und den Wirkungen hergestellt wurde. Hingegen können die Autoren zu Recht darauf hinweisen, dass die für alle drei Fälle geltenden Schlussfolgerungen besser abgestützt werden konnten, als wenn lediglich ein Fall in die Untersuchung einbezogen worden wäre. Damit sind auch die wesentlichen Vorteile genannt, die vergleichende Fallstudien gegenüber der Einzelfallstudie bieten. Sie bieten die Chance, unterschiedliche Wirkungszusammenhänge aufzudecken, und sie sichern die Befunde, die sich in allen untersuchten Fällen erhärten liessen, besser ab. Damit soll aber keinen Verallgemeinerungen über die untersuchten Fälle hinaus das Wort geredet werden. Fallstudien haben ihre Gültigkeit nur für die untersuchten Fälle, auch wenn sie vergleichend angelegt sind.

Damit soll aus einer empirisch-analytischen Position indessen nicht ausgeschlossen werden, dass mit Fallstudien keine Hypothesen überprüft werden können. In dieser Forschungstradition wird unterschieden zwischen Fallstudien, die eher der Hypothesenbildung dienen, und solchen, welche in erster Linie die Hypothesenprüfung vor Augen haben (vgl. u.a. Eckstein 1975). *Hypothesenbildende Fallstudien* haben die Aufgabe, komplexe Zusammenhänge aufzuhellen und so zur Entwicklung empirisch begründbarer theoretischer Konzepte, Theorien, Hypothesen beizutragen (Lamnek 1989: 11). Sie dienen somit dem Kennenlernen und der Interpre-

tation von Problemzusammenhängen und der Formulierung von empirisch relevanten Hypothesen. Diese lassen sich in der nachfolgenden Hauptuntersuchung dann testen. Bei hypothesenbildenden Fallstudien werden meist typische Fälle untersucht, d.h. solche, welche den "Normalfall" repräsentieren. *Hypothesenprüfende Fallstudien* dagegen analysieren eher kritische Fälle. Damit sollen Hypothesen bestätigt, relativiert oder verworfen werden. Ausgangspunkt bilden Annahmen, welche das Eintreten eines bestimmten Ereignisses gewissermassen zwingend voraussagen. Zeigt eine Fallstudie nun ein Ausbleiben des unter den angegebenen Ausgangsbedingungen prognostizierten Ereignisses auf, so ist eine Modifikation oder eine Verwerfung der verwendeten Hypothese notwendig. Anderenfalls gilt die Hypothese als vorläufig bestätigt. Ein Beispiel für hypothesentestende vergleichende Fallstudien aus der schweizerischen Evaluationsforschung findet sich bei Knoepfel, Imhof und Zimmermann (1995) im Rahmen der Evaluation der lufthygienischen Wirkungen der Massnahmenpläne.

11.2.3. Der quasi-experimentelle Vorher/Nachher-Vergleich

Vom Design der hypothesentestenden vergleichenden Fallstudien zum hier folgenden quasi-experimentellen Design des Vergleichs von Situationen "vorher" und "nachher" ist es nur noch ein kleiner Schritt. Das letzte Design geht von der Annahme aus, dass sich bei einem Fall, der einer staatlichen Massnahme ausgesetzt ist, verschiedene Merkmale anders entwickeln als bei dem anderen Fall, der nicht mit derselben Massnahme konfrontiert ist. Die Untersuchungsanlage ist denn auch relativ einfach. Man wähle mindestens zwei Fälle aus, die sich möglichst nur dadurch unterschieden, dass im einen Fall ein Programm entwickelt und umgesetzt wird, im anderen dagegen nicht. Daraufhin vergleicht man in beiden Fällen die Situation vor der Einführung der Massnahme mit der Situation nachher. Diese Vorgehensweise geht aus der Abbildung 5 hervor.

Abbildung 5: Vorher/Nachher-Vergleich

Fälle (Kanton A und B)	Situation (Arbeitslosenrate)	
	vorher	nachher
Fall A (mit Massnahme)	5%	4%
Fall B (ohne Massnahme)	5%	5%

Hat sich die Situation im Fall A mit Massnahme verbessert, im Fall B ohne Massnahme dagegen nicht, dann gilt die Massnahme - im erwünschten Sinne - als wirksam. Ist also etwa im Kanton A aufgrund eines Arbeitsbeschaffungsprogramms die Arbeitslosigkeit von 5 auf 4 Prozent gesunken, im Kanton B ohne Arbeitsbeschaffungsprogramm dagegen bei 5 Prozent geblieben, so kann vorerst einmal vermutet werden, dass das Programm im Kanton A erfolgreich war.

Selbstverständlich liegen die Dinge in der Praxis nie so einfach. Die wichtigsten Schwierigkeiten dieses Designs sind die folgenden:

- Erstens ist die Ausgangslage kaum je in den beiden Fällen die gleiche. So ist es wahrscheinlich, dass z.B. Kantone mit einer hohen Arbeitslosenrate eher ein Arbeitsbeschaffungsprogramm einführen als andere. Die Ausgangslage im Kanton A ist deshalb vermutlich schlechter als jene im Kanton B.

- Zweitens dürfte sich in der Regel in beiden Fällen die Situation verändern, wenn auch nicht in gleichem Ausmass. Die Arbeitslosenrate kann im Kanton A von 5 auf 4 Prozent, im Kanton B von 4,2 auf 3,6 Prozent sinken. Die Bestimmung des Ausmasses der Wirkung einer Massnahme wird so deutlich schwieriger.

- Drittens bleibt unklar, ob die Verbesserung der Situation im Kanton A etwas mit der getroffenen Massnahme zu tun hat. Der Vergleich der Situationen vorher und nachher gibt allein noch keinen Hinweis auf die Kausalität.

Um diese Schwierigkeiten etwas zu mindern, kann die Strategie befolgt werden, mehrere Fälle mit und ohne Programm einzubeziehen. Wenn alle

Fälle mit Programm eine Verbesserung der Situation in der Zieldimension erreichen, nicht aber die Fälle ohne Programm, dann besteht etwas bessere Gewähr für die Wirksamkeit der Massnahme. Allerdings handelt sich der Forscher oder die Forscherin damit ein neues Problem ein. Es ist nämlich unwahrscheinlich, dass in allen Fällen das Programm zum gleichen Zeitpunkt eingeführt wird und dass es überall gleich ausgestattet ist. Damit werden zwingend unterschiedliche Vorher/Nachher-Situationen miteinander verglichen. Eine saubere und eindeutige Untersuchungsanlage wäre nur bei einem reinen Experiment möglich. Dieses ist aber im Rahmen öffentlicher Politiken meist nicht durchführbar (vgl. Abschnitt 11.1.4.). Eine Verbesserung des Evaluationsdesigns kann zwar mit dem Einbezug einer grösseren Zahl von Fällen erreicht werden, weil so gewisse Variablen, die das Bild verfälschen, besser kontrolliert werden können. Diesem Vorgehen sind aber meist natürliche Grenzen gesetzt. So lässt sich etwa die Zahl der Kantone nicht vermehren.

Als *Beispiel* dient deshalb eine Evaluation, bei der die Fallzahl nicht a priori beschränkt war. In einer Vollzugs- und Wirkungsevaluation der Instrumente der *eidgenössischen Wohneigentumsförderung* haben Schulz, Muggli und Hübschle (1993) nicht nur das Verhalten der Vollzugsinstanzen und der Adressaten auf die Übereinstimmung mit den beschlossenen Massnahmen überprüft (Vollzugskontrolle), sondern auch untersucht, ob das Hauptziel der Wohneigentumsförderung, nämlich die Ermöglichung des Erwerbs von Grundeigentum durch Unterstützung bei der Finanzierung in der Form von Bürgschaften, erreicht worden ist (Wirksamkeitsprüfung). Zu diesem Zweck wurden verschiedene Untersuchungsgruppen gebildet und befragt. Die zentrale Gruppe bildeten die neuen Eigentümer, die eine Förderung erfuhren (N=188). Diesen wurden drei Kontrollgruppen gegenübergestellt. Der ersten Gruppe gehören jene an, die seit 1976 Wohneigentümer geworden waren, ohne von der Wohneigentumsförderung zu profitieren (N=262). Die zweite Gruppe wurde von Mietern ohne realisierte Kaufabsichten gebildet (N=385). Einer dritten, gemischten Gruppe gehörten Personen an, die ein Gesuch um Wohneigentumsförderung zurückgezogen hatten oder deren Gesuch abgelehnt worden war (N=83). Die Analyse anhand von Interviews zeigt, dass sich die Gruppe der Geförderten in mehrfacher Hinsicht signifikant von den anderen unterscheidet. Dennoch rechnen die Autoren damit, dass ein Drittel bis zur Hälfte der Geförderten zu den "Mitnehmern" gehörten, die die Immobilie auch ohne Wohneigentumsförderung gekauft hätten. Da zudem kaum je mehr als 10 Prozent des neu gebauten Wohneigentums der Förderung unterstand, darf auf Grund des Kontrollgruppenvergleichs angenommen werden, dass die Förderung des Wohneigentums ihr Ziel nur in einem sehr beschränkten Masse erreichte.

Das Beispiel zeigt eine zusätzliche Problematik des Vorher/Nachher-Vergleichs auf. Wenn die Situation im Zeitpunkt vor der Intervention nicht systematisch erfasst worden ist, wird es im Nachhinein äusserst schwierig, sie zu rekonstruieren. Hier wurde dies teilweise mit einer Befragung der Betroffenen versucht. Diesem Vorgehen sind freilich Grenzen gesetzt. Denn die Befragten können sich in der Regel nur schlecht an die frühere Lage erinnern. Ferner neigen sie dazu, diese Situation zu verklären, sei es in positiver oder negativer Richtung. Quasi-experimentelle Vorher/Nachher-Vergleiche eignen sich deshalb nur dann, wenn sich die Evaluation auf unproblematische Reihen von Aggregatdaten stützen kann oder wenn das Design vor der Durchführung der Massnahme geplant werden kann. Das ist etwa bei Schulversuchen (Einführung der Fünftagewoche, integrierte Oberstufe) der Fall. Eine solche Strategie wurde auch bei der Einführung von Tempo 50 innerorts in Testgemeinden gewählt (Arbeitsgruppe Verkehrssicherheit 1983). Rückwirkend ist es schwierig, das Vorher/Nachher-Design in reiner Form zu realisieren.

11.2.4. Die Zeitreihenanalyse

Die höchsten Anforderungen an die Datenqualität stellt die Zeitreihenanalyse. Hier geht man davon aus, dass sich die Wirksamkeit einer Massnahme daran zeigt, dass sich der Verlauf einer Reihe von Beobachtungen nach einer Intervention - über die üblichen zufälligen Schwankungen hinaus - gegenüber vorher verändert. Um trendbedingte sowie zufällige Entwicklungen in den Griff zu bekommen, beschränken sich entsprechende Untersuchungen nicht nur auf einen Vergleich der beiden Situationen vorher/nachher an zwei Stichdaten, sondern sie beobachten die Entwicklung kontinuierlicher während eines möglichst langen Zeitraums.

Soll die Zeitreihe mit statistischen Methoden analysiert werden, so sind je nach Auswertungsverfahren mindestens 20 - 50 Beobachtungen erforderlich.

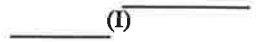
Schwierigkeiten bietet die Zeitreihenanalyse nicht zuletzt deshalb, weil in der Praxis die politisch-administrativen Interventionen nicht einmalig erfolgen. Vielmehr finden in der Regel mehrere relevante Interventionen

in nicht weit auseinanderliegenden Abständen statt. Es handelt sich um Prozesse mit sich überlagernden Einflüssen. Hinzu kommen zumeist exogene Faktoren, welche eine Zeitreihe ebenfalls beeinflussen. Die verschiedenen Wirkungen auf eine zeitliche Datenreihe analytisch auseinanderzuhalten, stellt höchste Anforderungen an die Qualität der Daten einerseits und an die Fähigkeiten der Forscherin oder des Forschers andererseits. Vielfach bleibt im Evaluationsalltag nur der pragmatische Weg, chronologische Entwicklungsverläufe einzelner Variablen in grafischen Darstellungen abzubilden und diese zu interpretieren.

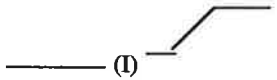
Das Untersuchungsdesign der Zeitreihe ist auch deshalb schwer zu handhaben, weil sehr verschiedene Wirkungsverläufe einer Intervention möglich sind und es nicht einfach ist, diese in klaren Hypothesen zu antizipieren. So kann eine Wirkung abrupt oder verzögert eintreten. Es kann auf der Zieldimension eine Verschiebung des Niveaus oder eine Veränderung der Richtung geben. Die Veränderung kann andauernd oder nur vorübergehend sein. Daraus lassen sich verschiedene Wirkungsverläufe kombinieren, wie sie im folgenden in Anlehnung an Glass, Willson und Gottman (1975) zusammengestellt werden.

Abbildung 6 Zeitpfade einer Wirkung (nach Glass, Willson und Gottman 1975: 44)

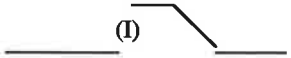
A. Abrupter Niveauwechsel



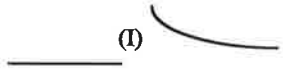
B. Verzögerter Niveauwechsel



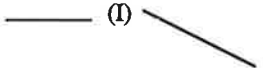
C. Vorübergehender Niveauwechsel



D. Verschwindender Niveauwechsel



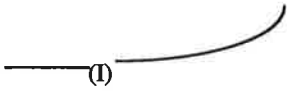
E. Abrupter Richtungswechsel



F. Verzögerter Richtungswechsel

G. Vorübergehender Richtungswechsel

H. Beschleunigter Richtungswechsel



I. Entwicklungseffekt

J. Veränderung der Variabilität

Ein *Beispiel* für eine statistisch gestützte Zeitreihenanalyse ist die Evaluation verschiedener Massnahmen der Luftreinhaltepolitik in der Schweiz von Thomas Widmer (1991). In dieser Studie versucht der Autor, den Wirkungsverlauf von mehreren Massnahmen zur Reduktion der Luftbelastung durch Schwefeldioxid anhand einer Interventionsanalyse nach dem Ansatz von Box und Tiao (1975) zu erfassen und zu interpretieren. Die Interventionsanalyse ergab, dass weder die Richtlinien aus den Jahren 1972 und 1983 noch verschiedene private Normen einen signifikanten Einfluss auf den Schwefelgehalt von Heizölen und Diesel-Treibstoff geschweige denn auf die Schwefeldioxid-Immissionen hatten. Einzig von der Luftreinhalte-Verordnung von 1985 ging eine Wirkung auf diese Zieldimensionen aus.

11.2.5. Kombination und Triangulation

Angesichts der Schwierigkeiten, die entstehen, wenn mit einem einzigen Evaluationsdesign die Wirkungen einer Massnahme präzise erfasst werden sollen, sind in den letzten Jahren wieder vermehrte Anstrengungen unternommen worden, mit Hilfe des kombinierten Einsatzes verschiedener Designs die Gültigkeit der Aussagen von Evaluationsstudien zu erhöhen. In diesem Zusammenhang hat auch die sogenannte *Triangulation* neue Aktualität gewonnen. Der Begriff stammt ursprünglich aus dem Bereich der militärischen Navigation und bezeichnet eine Strategie, um, ausgehend von verschiedenen Referenzpunkten, die exakte Position eines Objekts festzustellen. In die Sozialwissenschaften wurde der Gedanke von Campbell und Fiske (1959) eingeführt. Deren Idee war es, eine Hypothese mit einer Serie komplementärer Testmethoden zu überprüfen. Dabei sollte der Forscher eine Vermutung auf der Basis der ersten Methode durch Befunde einer zweiten Methode absichern. Zudem wurde die Chance in Betracht gezogen, dass bei der Anwendung der zweiten Technik auf überraschende Erkenntnisse gestossen werden könnte. Durch Methodentriangulation sollten weiter in den einzelnen Methoden steckende Fehler entdeckt werden. Es erschien als äusserst unwahrscheinlich, dass sich mit mehreren Techniken dieselben falschen Resultate ergeben würden.

In den siebziger Jahren wurde das Prinzip der Triangulation von Denzin (1989) präzisiert und ausdifferenziert. Er unterschied fünf Typen von Design-Triangulation:

Als *"Daten-Triangulation"* bezeichnete er den Einbezug verschiedener Datenquellen in eine Untersuchung. Unter *"Investigator-Triangulation"* verstand er den Einbezug unterschiedlicher Beobachter oder Forscher, um aktorbedingte Verzerrungen der Ergebnisse zu vermeiden. Die Kombination verschiedener Erhebungstechniken benannte er *"methodologische Triangulation"*. Den vierten Typ bezeichnete er als *"Theorien-Triangulation"*. Hierbei werden die Daten mit verschiedenen Theorien konfrontiert und so die Erklärungskraft der einzelnen Ansätze geprüft (vgl. u.a. Flick 1992). Als fünften Typus schliesslich fügte er die sogenannte *multiple Triangulation* hinzu. Dieser Typus beinhaltet die Idee, dass eine Triangulation der Datenquellen, der Forschungsteams, der Methoden und der Theorien gleichzeitig vorgenommen wird. Der Ansatz ist in höchstem Masse anspruchsvoll, verlangt eine eigentliche übergeordnete Forschungsstrategie und ist derart aufwendig, dass er für eine Evaluation nur ausnahmsweise in Frage kommen dürfte.

Auf andere Typen der Kombination von verschiedenen Forschungsansätzen kann die Evaluationsforschung hingegen sehr wohl zurückgreifen. Besonders Michael Patton (1990) hat auf ihre Chancen hingewiesen. Eine besondere Form der Triangulation kann auch in der *Kombination der Vergleichsebenen* liegen. Balthasar und Knöpfel (1994) bezeichnen sie als *"konzeptionelle Triangulation"*. Sie besteht darin, dass der gleiche Gegenstand, d.h. dasselbe Evaluandum mit verschiedenen Analysekonzepten angegangen wird. Dieses Vorgehen hat zwei Vorteile: Einmal können aus den verschiedenen Teiluntersuchungen Ergebnisse hervorgehen, welche als Fragestellungen, als Datengrundlage oder als Randbedingung in andere Teile einfliessen können. Dann ist zu erwarten, dass - ähnlich wie in der Fotografie - auch in der Evaluation zusätzliche "Lampen" einen Gegenstand besser auszuleuchten und unerwünschte Schatten zu beseitigen vermögen (vgl. dazu auch Balthasar und Knöpfel 1994: 166). Durch unterschiedliche Zugangsweisen wird die Abhängigkeit der Ergebnisse von einzelnen Analysekonzepten relativiert.

Siegfried Lamnek (1989) hebt als Vertreter eines alternativen Ansatzes allerdings hervor, dass Triangulationen keine "Hauptwirklichkeit" erkennen lassen, sondern zu komplementären Bildern der Realität führen. Triangulation sollte demnach in dieser Sicht weniger der Konstruktion einer "objektiven" Wahrheit dienen, als bewusst die Breite und die Tiefe der Analyse erweitern. Es sind von den einzelnen Untersuchungsteilen nicht deckungsgleiche, sondern komplementäre Ergebnisse zu erwarten, welche

sich ineinander fügen und sich ergänzen können, aber nicht kongruent sein müssen (Lamnek 1988: 236).

Als *Beispiel* für eine Studie mit mehrfachem Zugang dient uns die Evaluation von Balthasar und Knoepfel (1994). Sie befasst sich mit der Analyse der *innovationsfördernden Auswirkungen* eines Programms der schweizerischen Umweltpolitik. Drei *luftreinhaltepolitische Massnahmen* im Bereich der Hausfeuerungen werden daraufhin untersucht, ob sie zu technischen Innovationen führen, welche die geforderte Abgasqualität ermöglichen lassen. Die beiden Autoren bearbeiten die Thematik mit unterschiedlichen Vorgehensweisen, die sie als "Analysekonzepte" bezeichnen. Die Auswirkungen der Typenprüfung für Ölbrenner und Heizungskessel werden anhand einer Marktanalyse untersucht. Bei den Wirkungen der Feuerungskontrolle werden eine Umfrage und eine multiple Regression durchgeführt. Bei der Massnahme der Grenzwertsetzung stützt sich die Evaluation auf eine Fallstudie. Die Untersuchung des gesamten dreiteiligen Programms erfolgt über eine statistische Zeitreihenanalyse. Die verschiedenen Ansätze ergänzen sich gut und erlauben es, vertiefte Erkenntnisse über die Wirkungszusammenhänge zu gewinnen. Um eine "Triangulation" im eigentlichen Sinne handelt es sich allerdings nicht, da nur jede Teilmassnahme aufgrund eines eigenen Designs auf ihre Wirkungen untersucht, nicht aber ein einzelnes Evaluandum mit je verschiedenen Forschungsansätzen analysiert wird.

Dies war bei einem anderen *Beispiel* der Fall. Die "Arbeitsgruppe Gesetzesevaluation (AGEVAL)" des Eidgenössischen Justiz- und Polizeidepartements gab zwei Forschungsequipen den Auftrag, mit je einem anderen Untersuchungsdesign die Wirkungen der *beruflichen Vorsorge* auf den Arbeitsmarkt zu analysieren. Gerheuser (1991) packte die Aufgabe im Sinne einer Fallstudie an und stützte seine Analyse auf die Aussagen von Personalverantwortlichen von 42 Unternehmen und (halb-) öffentlichen Institutionen; die Gespräche fanden im Rahmen von elf Hearings in verschiedenen Regionen der Schweiz statt. Schaetti (1990) stützte sich dagegen auf Daten der Zentralen Ausgleichsstelle der AHV in Genf und "verglich die Häufigkeit und Wahrscheinlichkeit von Arbeitslosigkeit, Stellenwechseln und gering entlöhnten Arbeitsverhältnissen vor und nach der Einführung des Obligatoriums" ("Arbeitsgruppe Gesetzesevaluation" 1991) der beruflichen Vorsorge. Diese beiden Designs zur Analyse der Wirkungen derselben Massnahme ergänzten sich und brachten die erwünschten komplementären Ergebnisse. Das Vorgehen entspricht einer

gleichzeitigen Triangulation der "Methoden" und der "Investigatoren" und hat in diesem Sinne für die Schweiz Modellcharakter.

Aus diesen Beispielen ist freilich nicht der Schluss zu ziehen, Evaluationen seien nur dann sinnvoll und liessen nur dann gültige Schlüsse zu, wenn sie sich auf Studien stützen, denen eine Kombination von mehreren Untersuchungsdesigns zugrunde liegt. Ein solches Vorgehen nach dem Vorbild der Triangulation ist aufwendig und kostspielig; es sollte nur dann zum Einsatz gelangen, wenn viel auf dem Spiel steht und von der Evaluation weitreichende Entscheidungen abhängen. Bei weniger umstrittenen Massnahmen und bei Programmen, die mit einfachen Instrumenten arbeiten, genügen durchaus Untersuchungsdesigns, wie wir sie vorher in diesem Kapitel geschildert haben. In jedem Falle aber gilt, dass die Wahl des Evaluationsdesigns immer zuerst von der Fragestellung abhängig zu machen ist. Als zweites Kriterium kommen die zur Verfügung stehenden Mittel in Betracht. Die persönlichen Präferenzen der Forschungsteams sollten eine untergeordnete Rolle spielen.